An Overall Index for Comparing Hierarchical Clusterings

I. Morlini and S. Zani

Abstract In this paper we suggest a new index for measuring the distance between two hierarchical clusterings. This index can be decomposed into the contributions pertaining to each stage of the hierarchies. We show the relations of such components with the currently used criteria for comparing two partitions. We obtain a similarity index as the complement to one of the suggested distances and we propose its adjustment for agreement due to chance. We consider the extension of the proposed distance and similarity measures to more than two dendrograms and their use for the consensus of classification and variable selection in cluster analysis.

1 Introduction

In cluster analysis, one may be interested in comparing two or more hierarchical clusterings obtained for the same set of n objects. Indeed, different clusterings may be obtained by using different linkages, different distances or different sets of variables. In the literature the most popular measures have been proposed for comparing two partitions obtained by cutting the trees at a certain stage of the two hierarchical procedures (Rand (1971); Fowlkes and Mallows (1983); Hubert and Arabie (1985); Meila (2007); Youness and Saporta (2010)). Less attention has been devoted to the comparison of the global results of two hierarchical classifications, i.e. two dendrograms obtained for the same set of objects. Sokal and Rohlf (1962) have

I. Morlini (✉)
Department of Economics, University of Modena and Reggio Emilia, Via Berengario 51, 41100 Modena, Italy
e-mail: isabella.morlini@unimore.it

S. Zani
Department of Economics, University of Parma, Via Kennedy 6, 43100 Parma, Italy
e-mail: sergio.zani@unipr.it

W. Gaul et al. (eds.), *Challenges at the Interface of Data Analysis, Computer Science, and Optimization*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-24466-7_4, © Springer-Verlag Berlin Heidelberg 2012

introduced the so-called cophenetic correlation coefficient (see also Rohlf 1982 and Lapointe and Legendre 1995). Baker (1974) has proposed the rank correlation between stages where pairs of objects combine in the tree for measuring the similarity between two hierarchical clusterings. Reilly et al. (2005) have discussed the use of Cohen's kappa in studying the agreement between two classifications.

In this work we suggest a new index for measuring the dissimilarity between two hierarchical clusterings. This index is a distance and can be decomposed into the contributions pertaining to each stage of the hierarchies. In Sect. 2 we define the new index for two dendrograms. We then present its properties and its decomposition with reference to each stage. Section 3 shows the relations of each component of the index with the currently used criteria for comparing two partitions. Section 4 considers the similarity index obtained as the complement to one of the suggested distances and shows that its single components obtained at each stage of the hierarchies can be related to the measure B_k suggested by Fowlkes and Mallows (1983). This section also deals with the adjustment of the similarity index for agreement due to chance. Section 5 considers the extension of the overall index to more than two clusterings. Section 6 gives some concluding remarks.

2 The Index and Its Properties

Suppose we have two hierarchical clusterings of the same number of objects, n. Let us consider the $N = n(n-1)/2$ pairs of objects and let us define, for each non trivial partition in k groups ($k = 2, \ldots, n-1$), a binary variable X_k with values $x_{ik} = 1$ if objects in pair i ($i = 1, \ldots, N$) are classified in the same cluster in partition in k groups and $x_{ik} = 0$ otherwise. A binary $(N \times (n-2))$ matrix $\mathbf{X}_g$ for each clustering g ($g = 1, 2$) may be derived, in which the columns are the binary variables X_k. A global measure of dissimilarity between the two clusterings may be defined as follows:

$$Z = \frac{\| \mathbf{X}_1 - \mathbf{X}_2 \|}{\| \mathbf{X}_1 \| + \| \mathbf{X}_2 \|} \tag{1}$$

where $\| \mathbf{A} \| = \sum_i \sum_k \| a_{ik} \|$ is the L_1 norm of the matrix $\mathbf{A}$. In expression (1), since the matrices involved take only binary values, the L_1 norm is equal to the square of the L_2 norm.

Index Z has the following properties:

- It is bounded in [0,1].
- $Z = 0$ if and only if the two hierarchical clusterings are identical and $Z = 1$ when the two clusterings have the maximum degree of dissimilarity, that is when for each partition in k groups and for each i, objects in pair i are in the same group in clustering 1 and in two different groups in clustering 2 (or vice versa).
- It is a distance, since it satisfies the conditions of non negativity, identity, symmetry and triangular inequality (Zani (1986)).

- The complement to 1 of Z is a similarity measure, since it satisfies the conditions of non negativity, normalization and symmetry.
- It does not depend on the group labels since it refers to pairs of objects.
- It may be decomposed in $(n-2)$ parts related to each pair of partitions in k groups since:

$$Z = \sum_k Z_k = \sum_k \sum_i \frac{|x_{1ik} - x_{2ik}|}{\| \mathbf{X}_1 \| + \| \mathbf{X}_2 \|} \tag{2}$$

The plot of Z_k versus k shows the distance between the two clusterings at each stage of the procedure.

3 The Comparison of Two Partitions in k Groups

Let us consider the comparison between two partitions in k groups obtained at a certain stage of the hierarchical procedures. The measurement of agreement between two partitions of the same set of objects is a well-known problem in the classification literature and different approaches have been suggested (see, i.e., Brusco and Steinley 2008; Denoeud 2008). In order to highlight the relation of the suggested index with the ones proposed in the literature, we present the so-called matching matrix $M_k = [m_{fj}]$ where m_{fj} indicates the number of objects placed in cluster f $(f = 1, \ldots, k)$ according to the first partition and in cluster j $(j = 1, \ldots, k)$, according to the second partition (Table 1). Information in Table 1 can be collapsed in a (2×2) contingency table, showing the cluster membership of the object pairs in each of the two partitions (Table 2).

The number of pairs which are placed in the same cluster according to both partitions is

$$T_k = \sum_{f=1}^{k} \sum_{j=1}^{k} \binom{m_{fj}}{2} = \frac{1}{2} \left[\sum_{f=1}^{k} \sum_{j=1}^{k} m_{fj}^2 - n \right] \tag{3}$$

Table 1 Matching matrix M_k

	1	...	j	...	k	Total
1	m_{11}	...	...	...	m_{1k}	$m_{1.}$
2	m_{21}	...	...	...	m_{2k}	$m_{2.}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
f	...	...	m_{fj}	...	...	$m_{f.}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
k	m_{k1}	...	m_{kj}	...	m_{kk}	$m_{k.}$
Total	$m_{.1}$	...	$m_{.j}$	...	$m_{.k}$	n

Table 2 Contingency table of the cluster membership of the N object pairs

First clustering ($g = 1$)	Second clustering ($g = 2$)		Sum
	Pairs in the same cluster	Pairs in different clusters	
Pairs in the same cluster	T_k	$P_k - T_k$	P_k
Pairs in different clusters	$Q_k - T_k$	$U_k = N + T_k - P_k - Q_k$	$N - P_k$
Sum	Q_k	$N - Q_k$	$N = n(n-1)/2$

The counts of pairs joined in each partition are:

$$P_k = \sum_{f=1}^{k} \binom{m_{f.}}{2} = \frac{1}{2}\left[\sum_{f=1}^{k} m_{f.}^2 - n\right] \tag{4}$$

$$Q_k = \sum_{j=1}^{k} \binom{m_{.j}}{2} = \frac{1}{2}\left[\sum_{j=1}^{k} m_{.j}^2 - n\right] \tag{5}$$

The numerator of formula (2) with reference to the two partitions in k groups can be expressed as a function of the previous quantities:

$$\sum_{i=1}^{N} |x_{1ik} - x_{2ik}| = P_k + Q_k - 2T_k \tag{6}$$

The well-known Rand index (Rand 1971) computed for two partitions in k groups is given by (see Warrens 2008, for the derivation of the Rand index in terms of the quantities in Table 2):

$$R_k = \frac{N - P_k - Q_k + 2T_k}{N} \tag{7}$$

Therefore, the numerator of Z_k in (2) can be expressed as a function of the Rand index:

$$\sum_{i=1}^{N} |x_{1ik} - x_{2ik}| = N(R_k - 1) \tag{8}$$

The information in Table 2 can also be summarized by a similarity index, e.g. the simple matching coefficient (Sokal and Michener 1958):

$${}_{SM}I_k = \frac{T_k + (N + T_k - P_k - Q_k)}{N} = \frac{N + 2T_k - P_k - Q_k}{N} \tag{9}$$

If the Rand index is formulated in terms of the quantities in Table 2 it is equivalent to the simple matching coefficient and can be written as:

$$\sum_{i=1}^{N} |x_{1ik} - x_{2ik}| = N({}_{SM}I_k - 1) \tag{10}$$

4 The Complement of the Index

Since $\| \mathbf{X}_1 \| = \sum_k Q_k$ and $\| \mathbf{X}_2 \| = \sum_k P_k$, the complement to 1 of Z is:

$$S = 1 - Z = \frac{2 \sum_k T_k}{\sum_k Q_k + \sum_k P_k} \tag{11}$$

Also the similarity index S may be decomposed in $(n-2)$ parts V_k related to each pair of partitions in k groups:

$$S = \sum_k V_k = \sum_k \frac{2T_k}{\sum_k Q_k + \sum_k P_k} \tag{12}$$

The components V_k, however, are not similarity indices for each k since they assume values < 1 even if the two partitions in k groups are identical. For this reason, we consider the complement to 1 of each Z_k in order to obtain a single similarity index for each pair of partitions:

$$S_k = 1 - Z_k = \frac{\sum_{j=2}^{n-1} P_j + \sum_{j=2}^{n-1} Q_j - P_k - Q_k + 2T_k}{\sum_{j=2}^{n-1} P_j + \sum_{j=2}^{n-1} Q_j} \tag{13}$$

Expression (13) can be written as:

$$S_k = \frac{\sum_{j \neq k} P_j + \sum_{j \neq k} Q_j + 2T_k}{\sum_j P_j + \sum_j Q_j} \tag{14}$$

The index suggested by Fowlkes and Mallows (1983) for two partitions in k groups in our notation is given by:

$$B_k = \frac{2T_k}{\sqrt{2P_k 2Q_k}} = \frac{T_k}{\sqrt{P_k Q_k}} \tag{15}$$

The statistics B_k and S_k may be thought of as resulting from two different methods of scaling T_k to lie in the unit interval. Furthermore, in S_k and B_k the pairs U_k (see Table 2), which are not joined in either of the clusterings, are not considered as indicative of similarity. On the contrary, in the Rand index, the pairs U_k are considered as indicative of similarity. With many clusters, U_k must necessarily be large and the inclusion of this count makes R_k tending to 1, for large k. How the treatment of the pairs U_k may influence so much the values of R_k for different k or the values of R_k and B_k, for the same k, is illustrated in Wallace (1983).

A similarity index between two partitions may be adjusted for agreement due to chance (Hubert and Arabie 1985; Albatineh et al. 2006; Warrens 2008). With reference to formula (13) the adjusted similarity index AS_k has the form:

$$AS_k = \frac{S_k - E(S_k)}{max(S_k) - E(S_k)} \tag{16}$$

Under the hypothesis of independence of the two partitions, the expectation of T_k in Table 2 is:

$$E(T_k) = P_k Q_k / N \tag{17}$$

Therefore, the expectation of S_k is given by:

$$E(S_k) = \frac{\sum_{j \neq k} P_j + \sum_{j \neq k} Q_j + 2P_k Q_k / N}{\sum_j P_j + \sum_j Q_j} \tag{18}$$

Given that $max(S_k) = 1$, we obtain:

$$AS_k = \frac{\frac{\sum_{j \neq k} P_j + \sum_{j \neq k} Q_j + 2T_k - \sum_{j \neq k} P_j - \sum_{j \neq k} Q_j - 2P_k Q_k / N}{\sum_k P_k + \sum_k Q_k}}{\frac{\sum_k P_k + \sum_k Q_k - \sum_{j \neq k} P_j - \sum_{j \neq k} Q_j - 2P_k Q_k / N}{\sum_k P_k + \sum_k Q_k}} \tag{19}$$

Simplifying terms, this reduces to:

$$AS_k = \frac{2T_k - 2P_k Q_k / N}{P_k + Q_k - 2P_k Q_k / N} \tag{20}$$

The adjusted Rand index for two partitions in k groups is given by (Warrens 2008):

$$AR_k = \frac{2(NT_k - P_k Q_k)}{N(P_k + Q_k) - 2P_k Q_k} \tag{21}$$

and so AS_k is equal to the Adjusted Rand Index.

5 Extension to More than Two Clusterings

When a set of G ($G > 2$) hierarchical clusterings for the same set of objects is available, we may be interested to gain insights into the relations of the different classifications. The index Z defined in (1) may be applied to each pair of clusterings in order to produce a $G \times G$ distance matrix:

$$\mathbf{Z} = [Z_{gh}], \qquad g, h = 1, \ldots, G. \tag{22}$$

Furthermore, considering the index S defined in (11) for each pair of dendrograms, we obtain a $G \times G$ similarity matrix:

$$\mathbf{S} = [S_{gh}], \qquad g, h = 1, \ldots, G \tag{23}$$

that displays the proximities between each pair of classifications. Usually, the G clusterings are obtained applying different algorithms to the same data set. In this case, matrices $\mathbf{Z}$ and $\mathbf{S}$ may be useful in the context of the "consensus of classifications", i.e. the problem or reconciling clustering information coming from different methods (Gordon and Vichi 1998; Krieger and Green 1999). Clusterings with high distances (or low similarities) from all the others can be deleted before computing the single (consensus) clustering.

Indexes Z and S can also be used for variable selection in cluster analysis (Fowlkes et al. 1988; Fraiman et al. 2008; Steinley and Brusco 2008). The inclusion of "noisy" variables can actually degrade the ability of clustering procedures to recover the true underlying structure. For a set of p variables and a certain clustering method, we suggest different approaches.

First we may obtain the p one dimensional clustering with reference to each single variable and then compute the $p \times p$ similarity matrix $\mathbf{S}$. The pairs of variables reflecting the same underlying structure show high similarity and can be used to obtain a multidimensional classification. On the contrary, the noisy variables should present a similarity with the other variables near to the expected value for chance agreement. We may select a subset of variables that best explains the classification into homogeneous groups. These variables help us to better understand the multivariate structure and suggest a dimension reduction that can be used in a new data set for the same problem (Fraiman et al. 2008).

A second approach consists in finding the similarities between clusterings obtained with subsets of variables (regarding, for example, different features). This approach is helpful in finding aspects that lead to similar partitions and subsets of variables that, on the contrary, lead to different clusterings.

A third way to proceed consists in finding the similarities between the "master" clustering obtained by considering all variables and the clusterings obtained by eliminating each single variable in turn, in order to highlight the "marginal" contribution of each variable to the master structure.

6 Concluding Remarks

In this paper we have introduced a new index to compare two hierarchical clusterings. This measure is a distance and it is appealing since it does summarize the dissimilarity by one number and can be decomposed in contributions relative to each pair of partitions. This "additive" feature is necessary for comparisons with other indices and for interpretability purposes. The complement to 1 of the suggested measure is a similarity index and it also can be expressed a sum of the components with reference to each stage of the clustering procedure.

The new distance is a measure of dissimilarity of two sequences of partitions of n objects into $2, 3, \ldots, n-2, n-1$ groups. The fact that these partitions came from successive cutting of two hierarchical trees is irrelevant. The partitions could also

come from a sequence of non hierarchical clusterings (obtained, i.e., by k-means methods with a different number of groups).

Further studies are needed in order to illustrate the performance of the suggested indices on both real and simulated data sets.

References

Albatineh AN, Niewiadomska-Bugaj M, Mihalko D (2006) On similarity indices and correction for chance agreement. J Classification 23:301–313

Baker FB (1974) Stability of two hierarchical grouping techniques. Case I: Sensitivity to data errors. JASA 69:440–445

Brusco MJ, Steinley D (2008) A binary integer program to maximize the agreement between partitions. J Classification 25:185–193

Denoeud L (2008) Transfer distance between partitions. Adv Data Anal Classification 2:279–294

Fowlkes EB, Mallows CL (1983) A method for comparing two hierarchical clusterings. JASA 78:553–569

Fowlkes EB, Gnanadesikan R, Kettenring JR (1988) Variable selection in clustering. J Classification 5:205–228

Fraiman R, Justel A, Svarc M (2008) Selection of variables for cluster analysis and classification rules. JASA 103:1294–1303

Gordon AD, Vichi M (1998) Partitions of partitions. J Classification 15:265–285

Hubert LJ, Arabie P (1985) Comparing partitions. J Classification 2:193–218

Krieger AM, Green PE (1999) A generalized Rand-index method for consensus clusterings of separate partitions of the same data base. J Classification 16:63–89

Lapointe FJ, Legendre P (1995) Comparison tests for dendrograms: A comparative evaluation. J Classification 12:265–282

Meila M (2007) Comparing clusterings – an information based distance. J Multivariate Anal 98(5):873–895

Rand WM (1971) Objective criteria for the evaluation of clustering methods. JASA 66:846–850

Reilly C, Wang C, Rutherford M (2005) A rapid method for the comparison of cluster analyses. Statistica Sinica 15:19–33

Rohlf FJ (1982) Consensus indices for comparing classifications. Math Biosci 59:131–144

Sokal RR, Michener CD (1958) A statistical method for evaluating systematic relationships. Univ Kans Sci Bull 38:1409–1438

Sokal RR, Rohlf FJ (1962) The comparison for dendrograms by objective methods. Taxon 11:33–40

Steinley D, Brusco MJ (2008) Selection of variables in cluster analysis: An empirical comparison of eight procedures. Psychometrika 73:125–144

Wallace DL (1983) Comment on the paper "A method for comparing two hierarchical clusterings". JASA 78:569–578

Warrens MJ (2008) On the equivalence of Cohen's Kappa and the Hubert-Arabie adjusted Rand index. J Classification 25:177–183

Youness G, Saporta G (2010) Comparing partitions of two sets of units based on the same variables. Adv Data Anal Classification 4,1:53–64

Zani S (1986) Some measures for the comparison of data matrices. In: Proceedings of the XXXIII Meeting of the Italian Statistical Society, Bari, Italy, pp 157–169